WORDNET: A LEXICAL DATABASE FOR ENGLISH

George A. Miller, Principal Investigator

Cognitive Science Laboratory
Princeton University
Princeton, New Jersey 08542

PROJECT GOALS

The goal of this project is to provide lexical resources for natural language research. The primary emphases are on the further development and dissemination of the on-line lexical database, WordNet. A secondary goal is to learn how to develop contextual representations for different senses of a polysemous word, where a contextual representation is comprised of topical and local context for each sense.

RECENT RESULTS

WordNet Upgrade: The basic unit in WordNet is the synonym set, or synset, which represents a lexicalized concept. Synsets are comprised of open class words (nouns, verbs, adjectives, and adverbs) and are connected by bi-directional pointers denoting such semantic relations as synonymy, antonymy, hyponymy, meronymy, troponymy, entailment, etc. WordNet has been growing at a rate of about 1,500 lexicalized concepts (synsets) per month. The major source of new entries has been from the COMLEX vocabulary adopted by Ralph Grishman at NYU. As of February 9, 1994, WordNet contains 81,658 synsets and 109,570 unique character strings (words and collocations).

Semantic Concordance: A semantic concordance is a textual corpus and a lexicon so combined that every substantive word in the text is linked to its appropriate sense in the lexicon. In this case, the textual corpus is the Brown Corpus and the lexicon is WordNet. 103 passages from the Brown Corpus have been tagged with WordNet version 1.4 senses, and is available to authorized users of the Brown Corpus. For information, send email to wordnet@princeton.edu.

ESCORT, an X Windows interface to the semantic concordance, has been developed. ESCORT allows a user to search the semantic concordance for sentences containing words that have been semantically tagged. In addition, ESCORT has been incorporated into the WordNet interface so that when a user looks up a word, the option of seeing sentences that use the word in a given sense is available.

With support from the Linguistic Data Consortium, a semantic concordance is being prepared for the most frequent polysemous words from the COMLEX vocabulary.

WordNet Distribution: WordNet version 1.4 was released in August, a new and larger version of the lexical database with support for Sun 3, Sun 4, NeXT, Silicon Graphics, DECstation, IBM RS-6000, PC, and MacIntosh machines, and with upgraded documentation. The WordNet database is available via anonymous ftp from clarity.princeton.edu.

As of February, 1994, we have 280 registered users of WordNet 1.3 or 1.4 in 25 countries. In April, a WordNet users group was created and now has 148 members. For information, send email to wordnet@princeton.edu.

Word Sense Identification: Benchmarks for the performance of automatic sense-identification systems have been developed, based on statistics derived from the semantic concordance for the Brown Corpus. In addition, a corpus of sentences using the verb "serve" in different senses has been compiled and used, along with a similar corpus of the noun "line" to find topical and local context for word sense identification. In collaboration with staff members from Siemens Corporate Research, the use of local context was found to improve performance in identifying the intended sense of those polysemous words. In addition, a method was developed using similarity measures based on the structure of WordNet that increased the virtual size of training sets.

PLANS FOR THE COMING YEAR

We will continue to maintain, edit, and upgrade WordNet and make it available to interested users. When the COMLEX vocabulary has been completely incorporated into the WordNet database, we plan to release version 1.5 and make it available on a CD-ROM. In addition, the semantic concordance will continue to grow, and will also be made available via ftp.

In collaboration with colleagues from Siemens Corporate Research and Hunter College, research on word sense identification will continue.